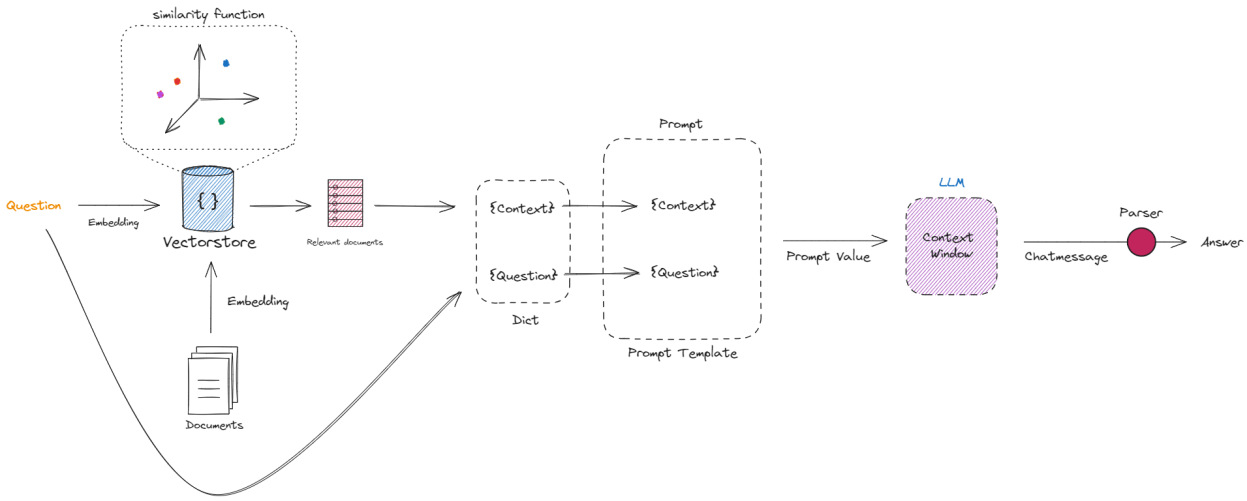
RAG deployment tutorial



In this series, we want to guide you through the a basic deployment of a LLM application.

▼ Introduction to LLM

- Transformer architecture
- encoding techniques
- What are prompts and prompt templates?
- What are the parameters that can be tuned?
- What are context windows and how do I modify LLM outputs?
- Multi-modal models

▼ Core Components

▼ Frontend

- Vercel's AI sdk
- SvelteKit

▼ Vector database

- What is a vector?

- Why vectordb than usual db? How about just elasticsearch?
- How does it work?
- What kind of encoding?

▼ Orchestration tools

- Langchain

▼ LLM Inference tool

- Model-As-a-Service
  - OpenAI
  - Azure OpenAI
  - Anthropic Claude
  - Google Vertex AI
- Self-hosted
  - Ollama

▼ Hosting

Fly.io

azure

Google

▼ Observability and Monitoring

▼ Implementation Variances

- Kubernetes based implemenetation

▼ Retrain and Finetuning

How to do it?

▼ Scaling Vectors

How and in what ways will your LLM apps scale?

- Adding different type of data that needs to be indexed and vectorized in your vectordb
- Reliability and data amount in your vectorDB
- QPS and speed of return per request, availability
- Security and Data privacy concerns
- Support for multi-modal LLMs
- Regressive behavior of models + benchmarking finetune results
- Monitor for misuse

▼ Helpful links

- https://artificialanalysis.ai/

  An overview of pricing, speed, and quality over llm API providers

-